Adrien Damez, Gavriil Kharkhordine, Elena Perotti, Angelo Picerno, Arnaud Van de Velde

P15 - Hawkes Processes

Reproduction and Application of an Asymptotically Distribution-Free Goodness-of-Fit Test for Hawkes Processes on High-Frequency Financial Data

January 21, 2026

Abstract

We study goodness-of-fit testing for temporal point processes using the asymptotically distribution-free procedure proposed by Baars, Can, and Laeven (2025). Focusing on univariate Hawkes processes, we first reproduce the main simulation results of the original paper under several parametric specifications. We then extend the analysis by conducting a detailed empirical study on high-frequency financial data, where model parameters must be estimated and standard assumptions are only approximately satisfied. By comparing the transformation-based procedure to naive goodness-of-fit diagnostics, including tests based on the random time change theorem, we show that naive approaches can be misleading in practice, either being overly conservative or rejecting almost systematically once the fitted model departs from the data. Overall, our results highlight the importance of accounting for parameter estimation when assessing the adequacy of Hawkes models, and demonstrate the practical relevance of estimation-aware goodness-of-fit tests in empirical applications.

KEYWORDS: Hawkes processes, goodness-of-fit testing, point processes, random time change, martingale transformation, simulations.

1. Introduction

Hawkes processes are widely used to model self-exciting phenomena in a variety of applied fields, including high-frequency finance, seismology, epidemiology, and social dynamics. Their ability to capture clustering and feedback effects has made them a standard modeling tool for temporal point process data. Assessing the adequacy of a fitted Hawkes model is therefore a fundamental statistical problem.

Classical goodness-of-fit procedures for point processes are typically based on the random time change theorem, which states that, under the true model, the time-rescaled event sequence should follow a unit-rate Poisson process. In practice, however, model parameters are unknown and must be estimated from data. More generally, parameter estimation uncertainty is often ignored when applying time-rescaling-based goodness-of-fit tests, despite the fact that this practice may substantially distort their finite-sample and asymptotic behavior.

Recently, Baars et al. [2025] proposed an asymptotically distribution-free goodness-of-fit testing procedure for general temporal point processes that explicitly accounts for parameter estimation uncertainty. By applying an innovation martingale transformation to a compensated empirical process, their method yields test statistics whose asymptotic distribution is independent of both the underlying model and the true parameter value. This provides a theoretically sound alternative to classical goodness-of-fit tests based on naive time rescaling.

The objective of this report is twofold. First, we reproduce the main numerical results of Baars et al. [2025] in the specific context of univariate Hawkes processes with exponential, power-law, and multi-exponential kernels. Second, we assess the practical performance of the transformation-based procedure through additional simulation experiments and an application to real-world high-frequency data. In particular, we explicitly compare the proposed test to commonly used random-time-change-based goodness-of-fit procedures and highlight the limitations of ignoring parameter estimation uncertainty in empirical practice.

2. Hawkes Process Models

We consider univariate linear Hawkes processes defined on the time interval $[0, T]$. A Hawkes process is a self-exciting point process, meaning that the occurrence of an event increases the likelihood of future events occurring shortly thereafter. Such models are widely used to describe clustered event arrivals, for instance in high-frequency financial markets where trades tend to arrive in bursts rather than at a constant rate.

Let N denote a simple point process with event times $\{t_i\}_{i\geq 1}$ and associated counting process $N(t) = \sum_i \mathbf{1}_{\{t_i \leq t\}}$, which counts the number of events that have occurred up to time t. The process is characterized by its conditional intensity function $\lambda_\theta(t)$, which plays the role of a local activity rate. More precisely, the conditional intensity is defined as

$$\lambda_\theta(t) = \lim_{\Delta t \to 0} \frac{1}{\Delta t} \mathbb{E}[N(t + \Delta t) - N(t) \mid \mathcal{F}_t],$$

where $\mathcal{F}_t$ denotes the information generated by the process up to time t. In this sense, $\lambda_\theta(t)$ can be interpreted as the expected number of events per unit of time at time t, given the past history of the process.

For a linear Hawkes process, the conditional intensity takes the form

$$\lambda_\theta(t) = \mu + \sum_{t_i<t} g_\theta(t - t_i),$$

where $\mu > 0$ is the baseline intensity, representing the average event rate in the absence of past activity. Throughout this work, the baseline intensity μ is assumed to be constant. While this restricts the generality of the model, it is required for the theoretical validity of the goodness-of-fit procedure of Baars et al. [2025] and leads to a simpler and more stable estimation framework.

The function $g_\theta : \mathbb{R}_+ \to \mathbb{R}_+$ is a non-negative excitation kernel parameterized by θ. Each past event contributes additively to the current intensity, with an influence that depends on how long ago the event occurred. The shape of the kernel g_θ therefore controls both the strength and the persistence of the self-excitation effect.

We restrict attention to stationary Hawkes processes, which requires $\int_0^\infty g_\theta(s)\, ds < 1$. Intuitively, this condition ensures that the additional activity generated by any given event eventually dies out, so that the process admits a well-defined long-run average intensity.

In this work, we focus on a small set of commonly used parametric kernels. This restriction is motivated both by theoretical considerations and by their widespread use in empirical applications, particularly in financial microstructure. While the goodness-of-fit procedures studied in this paper apply to general parametric Hawkes models, we limit our analysis to the following specifications.

Exponential kernel. The exponential Hawkes process is defined by the kernel

$$g_\theta(t) = \alpha e^{-\beta t}, \qquad \alpha > 0,\ \beta > 0,$$

where $\theta = (\mu, \alpha, \beta)$. This model is widely used due to its analytical tractability and Markovian structure. The stationarity condition reduces to $\alpha/\beta < 1$.

Power-law kernel. To allow for long-range dependence and slower decay of excitation, we also consider a Hawkes process with a power-law kernel of the form

$$g_\theta(t) = \alpha(1 + t)^{-(\beta+1)}, \qquad \alpha > 0,\ \beta > 0.$$

Compared to the exponential specification, this model captures persistent self-excitation and has been shown to provide a better fit in situations where clustering effects decay slowly over time. The stationarity condition coincides with that of the exponential kernel.

Multi-exponential kernel. Finally, we consider a multi-exponential Hawkes process defined by

$$g_\theta(t) = \sum_{j=1}^{J} \alpha_j e^{-\beta_j t}, \qquad \alpha_j > 0,\ \beta_j > 0.$$

This specification can be interpreted as a flexible approximation of a general kernel by a finite mixture of exponentials. The stationarity requires $\sum_{j=1}^{J} \alpha_j/\beta_j < 1$.

To simulate sample paths of Hawkes processes, we rely on constructive representations of point processes defined through their conditional intensity. Given an intensity of the form

$$\lambda(t) = \mu + \sum_{t_i<t} g(t - t_i),$$

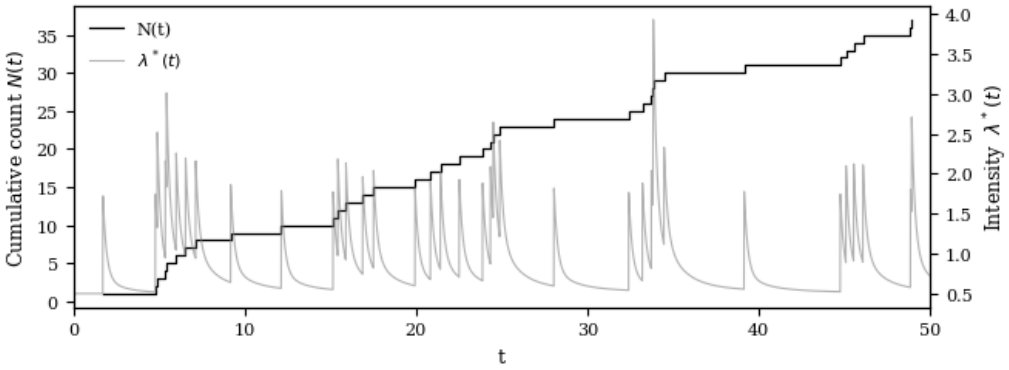

Figure 1. Hawkes process with a multi-exponential kernel: cumulative count $N(t)$ (black) and fitted intensity $\lambda^\star(t)$ (grey).

the goal is to generate event times sequentially on a finite time horizon $[0, T]$ in a way that is consistent with this stochastic intensity.

Two equivalent simulation principles are commonly used. The first is based on a branching (or cluster) representation of the process, while the second relies on Poisson thinning. Although both approaches generate sample paths with the same distribution, they emphasize different structural aspects of the process.

We first describe the branching representation, also known as the immigration–birth construction [Hawkes and Oakes, 1974].

In this representation, events are generated in two layers. First, immigrant events occur according to a homogeneous Poisson process with intensity μ, representing exogenous shocks to the system. Each immigrant then independently generates new events in the future.

Let g denote the excitation kernel and define the branching ratio

$$n = \int_0^\infty g(s)\, ds.$$

This quantity represents the expected number of direct offspring generated by a single event. Conditionally on an event occurring at time t, the number of its offspring is Poisson distributed with mean n, and their occurrence times are distributed according to the normalized kernel $g(s)/n$, shifted by t. Each offspring event subsequently generates further events in the same manner, giving rise to a branching cascade.

Under the stability condition $n < 1$, the expected total number of descendants of any immigrant is finite, and the superposition of all immigrants and their descendants yields a well-defined Hawkes process on the observation window $[0, T]$.

Alternatively, sample paths can be generated sequentially using Ogata's thinning method [Ogata, 1981]. In this approach, candidate events are proposed from a dominating Poisson process, and each candidate is accepted with probability equal to the ratio of the current intensity $\lambda(t)$ to the dominating rate. This method does not rely on an explicit branching interpretation.

Both approaches are implemented in our code. In the simulation study reported below, we generate sample paths using the cluster representation, while Poisson thinning is included for completeness and verification purposes.

In the simulation study, we specify parameter values for the Hawkes models introduced above. For the exponential and power-law kernels, we use the common parameter values

$$(\mu, \alpha, \beta) = (0.5,\ 1.0,\ 2.0).$$

For the multi-exponential specification, we take $J = 3$ components with

$$\mu = 0.5, \qquad (\alpha_0, \alpha_1, \alpha_2) = (1.0,\ 0.2,\ 0.02), \qquad (\beta_0, \beta_1, \beta_2) = (5.0,\ 1.0,\ 0.2).$$

The parameter values used in the simulation study are chosen such that all models have the same stationary mean intensity. For a stationary linear Hawkes process, the unconditional intensity satisfies

$$\bar{\lambda} = \frac{\mu}{1 - \int_0^\infty g_\theta(s)\, ds}.$$

For the exponential and power-law kernels considered above, the branching ratio $\int_0^\infty g_\theta(s)\, ds$ reduces to α/β. For the multi-exponential specification, it is given by $\sum_{j=1}^{J} \alpha_j/\beta_j$.

This normalization ensures that all models generate, one event per unit of time on average, so that differences observed in the simulation results can be attributed to the excitation structure rather than to differences in overall event rates.

Across all models, parameters are estimated by maximum likelihood based on the observation of N over $[0, T]$.

3. Parameter Estimation via Maximum Likelihood

Model parameters are estimated by maximum likelihood based on an observation of the point process N over the time interval $[0, T]$. Let $\{t_i\}_{i=1}^{n(T)}$ denote the observed event times up to time T. For a univariate Hawkes process with conditional intensity $\lambda_\theta(t)$, the log-likelihood function is given by

$$\ell_T(\theta) = \sum_{i=1}^{n(T)} \log \lambda_\theta(t_i) - \int_0^T \lambda_\theta(t)\, dt.$$

The maximum likelihood estimator $\hat{\theta}_T$ is defined as any maximizer of $\ell_T(\theta)$ over the admissible parameter space, subject to the positivity and stationarity constraints imposed by the model.

In this section, we briefly describe the explicit form of the likelihood under the different kernel specifications considered in this work, and discuss the associated computational aspects.

Exponential kernel. For the exponential Hawkes process with kernel $g_\theta(t) = \alpha e^{-\beta t}$, the conditional intensity takes the form

$$\lambda_\theta(t) = \mu + \alpha \sum_{t_i < t} e^{-\beta(t - t_i)}.$$

In this case, the likelihood admits an efficient recursive evaluation. Defining

$$R(t_i) = \sum_{t_j < t_i} e^{-\beta(t_i - t_j)},$$

one obtains the recursion

$$R(t_i) = e^{-\beta(t_i - t_{i-1})}(1 + R(t_{i-1})),$$

which allows the intensity $\lambda_\theta(t_i)$ to be computed in constant time per event. As a result, the evaluation of the log-likelihood scales linearly with the number of observed events. This computational efficiency, together with the Markovian structure of the model, makes the exponential kernel particularly attractive in large-scale applications.

Power-law kernel. For the power-law specification

$$g_\theta(t) = \alpha(1 + t)^{-(\beta+1)},$$

the conditional intensity becomes

$$\lambda_\theta(t) = \mu + \alpha \sum_{t_i < t} (1 + t - t_i)^{-(\beta+1)}.$$

Unlike the exponential case, no exact recursive representation is available for this kernel. The evaluation of the intensity at each event time requires summing over all past events, leading to a computational cost that grows quadratically with the number of observations. This reflects the long-memory nature of the power-law kernel, which captures persistent excitation effects at the expense of increased computational complexity.

Multi-exponential kernel. For the multi-exponential Hawkes process defined by

$$g_\theta(t) = \sum_{j=1}^{J} \alpha_j e^{-\beta_j t},$$

the intensity can be written as

$$\lambda_\theta(t) = \mu + \sum_{j=1}^{J} \alpha_j \sum_{t_i < t} e^{-\beta_j (t - t_i)}.$$

Each exponential component admits a recursion analogous to the single exponential case. Introducing

$$R_j(t_i) = \sum_{t_k < t_i} e^{-\beta_j (t_i - t_k)},$$

we have

$$R_j(t_i) = e^{-\beta_j (t_i - t_{i-1})}(1 + R_j(t_{i-1})), \qquad j = 1, \ldots, J.$$

The likelihood can therefore be evaluated in $\mathcal{O}(Jn(T))$ operations, which remains efficient as long as the number of components J is kept small. In practice, this specification provides a flexible compromise between computational tractability and the ability to approximate more general kernel shapes.

In all cases, parameters are estimated by numerical maximization of the log-likelihood under appropriate constraints. The resulting estimator $\hat{\theta}_T$ is used to construct fitted intensities and serves as input for the goodness-of-fit procedures described in the next section. As discussed below, treating $\hat{\theta}_T$ as fixed and ignoring the uncertainty induced by parameter estimation can lead to biased goodness-of-fit assessments and distorted test behavior.

4. Goodness-of-Fit Testing Procedure

Naive goodness-of-fit procedure. Let N be a univariate point process observed on the interval $[0, T]$, and assume that under the null hypothesis it belongs to a parametric family $\{N_\theta : \theta \in \Theta\}$ with conditional intensity λ_θ. Let $\hat{\theta}_T$ denote a parameter estimate obtained from the data. We first construct the compensated empirical process

$$\eta_T(u) = \frac{1}{\sqrt{T}} \left(N(uT) - \int_0^{uT} \lambda_{\hat{\theta}_T}(s)\, ds \right), \qquad u \in [0, 1].$$

Following Baars et al. [2025], we denote by $\mu_{\hat{\theta}_T} \approx \frac{N(T)}{T}$. the stationary mean intensity of the fitted model, that is, the quantity obtained by evaluating the unconditional mean intensity at the estimated parameter $\hat{\theta}_T$.

A naive goodness-of-fit procedure consists of standardizing this process and treating it as a Brownian motion. Specifically, we define

$$\tilde{W}_T(u) = \frac{1}{\sqrt{\mu_{\hat{\theta}_T}}}\, \eta_T(u),$$

The process $\tilde{W}_T$ is then treated as a standard Brownian motion, and goodness-of-fit tests are performed by applying classical functionals (e.g., Kolmogorov–Smirnov, Cramér–von Mises, or Anderson–Darling tests) to its increments. This procedure ignores the effect of parameter estimation in the compensator.

Transformation-based goodness-of-fit procedure. To obtain a correctly sized goodness-of-fit test, we apply a transformation to the compensated empirical process that accounts for parameter estimation. The transformed process is defined as

$$\widehat{W}_T(u) = \frac{1}{\sqrt{\mu_{\hat{\theta}_T}}} \left(\eta_T(u) - \int_0^u \frac{\eta_T(1) - \eta_T(v)}{1 - v}\, dv \right), \qquad u \in [0, 1].$$

In practice, goodness-of-fit testing is performed by selecting a fixed $\tau \in (0, 1)$ and an integer $n \geq 1$, and computing the normalized increments

$$\widehat{Z}_i^{(T)} = \sqrt{\frac{n}{\tau}} \Big(\widehat{W}_T(i\tau/n) - \widehat{W}_T((i-1)\tau/n) \Big), \qquad i = 1, \ldots, n.$$

Under the null hypothesis, the variables $(\widehat{Z}_i^{(T)})_{i=1}^n$ are asymptotically independent and identically distributed standard normal random variables. We therefore perform a normality test (such as Kolmogorov–Smirnov, Cramér–von Mises, or Anderson–Darling) on the sample $(\widehat{Z}_i^{(T)})_{i=1}^n$ to assess the goodness of fit of the model.

Naive random time change procedure. A commonly used goodness-of-fit approach for point processes is based on the random time change theorem. After estimating the parameter $\hat{\theta}_T$, one computes the compensator

$$\Lambda_{\hat{\theta}_T}(t) = \int_0^t \lambda_{\hat{\theta}_T}(s)\, ds,$$

and transforms the observed event times $(t_i)_{i \geq 1}$ into

$$\tau_i = \Lambda_{\hat{\theta}_T}(t_i).$$

Equivalently, one may consider the transformed interarrival times

$$x_i = \tau_i - \tau_{i-1} = \Lambda_{\hat{\theta}_T}(t_i) - \Lambda_{\hat{\theta}_T}(t_{i-1}), \qquad i \geq 2.$$

Under the random time change theorem, if the true parameter θ_0 were known, the variables (x_i) would form an i.i.d. sample from the standard exponential distribution. In practice, however, the true parameter is replaced by its estimator $\hat{\theta}_T$, and the resulting estimation error is ignored. Consequently, this procedure does not yield an asymptotically distribution-free test, even under the null hypothesis. Despite this theoretical limitation, random time change–based tests are widely used in empirical applications due to their simplicity and ease of implementation. For this reason, and in order to facilitate comparison with standard practices in the literature, we also report results obtained from this naive random time change procedure.

5. Simulation Study

This section presents a Monte Carlo simulation study designed to reproduce and validate the finite-sample behavior of the goodness-of-fit testing procedures introduced in Baars et al. [2025]. Our primary objective is to replicate the numerical results reported in that paper under parametric Hawkes null hypotheses, using the same model specifications, parameter values, and testing setup.

In addition to reproducing the original results, we extend the simulation study by including a naive goodness-of-fit procedure based on the random time change theorem. Although this procedure is discussed theoretically in Baars et al. [2025], it is not included in their simulation tables. Its inclusion allows us to compare different naive approaches that ignore parameter estimation uncertainty, and to assess how the use of the random time change transformation affects the finite-sample behavior of the tests.

We consider univariate Hawkes processes with exponential, power-law, and multi-exponential kernels, whose realizations are denoted by N^{ExpH}, N^{PLH}, and N^{MEH}, respectively. For each model, independent sample paths are simulated over a fixed observation window, after which model parameters are estimated via maximum likelihood. Goodness-of-fit is assessed using classical Kolmogorov–Smirnov, Cramér–von Mises, and Anderson–Darling tests applied to standardized increments constructed from transformed processes.

Tables 1 and 2 show that we successfully reproduce the main findings of Baars et al. [2025]. When the data-generating model coincides with the null hypothesis, the transformation-based procedure is correctly sized, with rejection frequencies close to the nominal levels, whereas the naive procedures that ignore estimation uncertainty are overly conservative. A striking additional feature is the behavior of the naive random-time-change (RTC) test: while it is undersized under the true null, it starts rejecting much more aggressively as soon as the fitted model deviates from the data-generating mechanism. This sensitivity to model misspecification, already visible in the simulation results, anticipates the systematic over-rejection of the RTC procedure observed later on real data.

Tables 3 and 4 extend the simulation study to the multi-exponential Hawkes setting, which was not considered in Baars et al. [2025]. Table 3 follows the same design as the previous experiments and considers the fully parametric multi-exponential model, where all kernel parameters are jointly estimated by maximum likelihood. The results exhibit

Table 1. Using the parametric null hypothesis H_0^{Exp}, with $T = 5{,}000$, $n = \lceil \sqrt{T}/4 \rceil$ and $\tau = 1$, we conduct 500 simulations for three different point process models. We report the number of rejections $R_{0.01}; R_{0.05}; R_{0.20}$ out of 500 tests, using significance levels 0.01, 0.05, and 0.20. Results are shown for three goodness-of-fit procedures described in Section 4: the transformation-based procedure, the naive goodness-of-fit procedure based on the compensated empirical process, and the naive random time change (RTC) procedure. Kolmogorov–Smirnov (KS), Cramér–von Mises (CvM), and Anderson–Darling (AD) tests are applied as described in Section 4.

	Transformation-based			Naive			Naive RTC		
Test	KS	CvM	AD	KS	CvM	AD	KS	CvM	AD
N^{ExpH}	4; 29; 96	4; 25; 99	4; 25; 93	0; 1; 6	0; 0; 3	0; 1; 5	0; 0; 1	0; 0; 0	0; 0; 13
N^{PLH}	5; 33; 111	5; 33; 108	11; 38; 129	0; 0; 8	0; 0; 7	0; 1; 23	0; 5; 69	0; 2; 33	16; 64; 262
N^{MEH}	12; 47; 130	14; 57; 118	28; 77; 174	0; 0; 25	0; 1; 21	0; 7; 48	2; 41; 204	0; 9; 199	131; 298; 463

Table 2. We report the number of rejections $R_{0.01}; R_{0.05}; R_{0.20}$ out of 500, using significance levels 0.01, 0.05, and 0.20 in an experiment analogous to the one conducted for Table 1, but now under the parametric null hypothesis H_0^{PL} with $T = 5{,}000$.

	Transformation-based			Naive			Naive RTC		
Test	KS	CvM	AD	KS	CvM	AD	KS	CvM	AD
N^{PLH}	7; 19; 101	8; 13; 111	8; 15; 113	0; 0; 6	0; 0; 4	0; 0; 7	0; 0; 1	0; 0; 0	0; 0; 13
N^{ExpH}	8; 23; 90	6; 23; 79	4; 20; 75	0; 0; 9	0; 0; 8	0; 0; 6	1; 5; 90	0; 1; 77	38; 124; 333
N^{MEH}	9; 29; 116	10; 33; 114	12; 46; 136	0; 0; 15	0; 0; 9	1; 4; 26	1; 2; 8	1; 2; 7	4; 17; 43

the same qualitative behavior as before: when the null hypothesis is correctly specified, the transformation-based procedure is well sized, whereas the naive procedures remain conservative. Moreover, as in the earlier tables, the naive RTC test starts rejecting much more frequently as soon as the fitted model deviates from the data-generating mechanism. In the multi-exponential setting, this behavior can already be observed under the true null hypothesis N^{MEH} and is likely driven by instability in the fitted intensity induced by the non-convexity of the estimation problem. Such local estimation errors have little impact on the transformation-based procedure, but are strongly amplified by the RTC transformation.

To address this issue, we consider the alternative estimation strategy reported in Table 4, where the decay parameters of the kernel are pre-estimated and subsequently held fixed, while only the remaining parameters are optimized. Specifically, the number of exponential components is fixed to three a priori, and the corresponding decay parameters are subsequently estimated based on a preliminary simulation experiment consisting of several independent sample paths. The precise procedure used to determine and fix these decay parameters is described in detail in Section 6. This setting allows us to assess whether the qualitative conclusions of the simulation study persist once estimation instability is reduced, and is the one adopted later in the real-data analysis. In this case, although estimation instability is mitigated, the naive RTC procedure may still reject aggressively when a multi-exponential kernel with fixed decay parameters is fitted to data generated from a simple exponential Hawkes process N^{ExpH}. Even though the decay parameters are calibrated on

exponential data, fitting a multi-scale kernel structure to a single-scale process requires delicate cancellations between excitation components, which can induce small local distortions in the compensator that are magnified by the RTC transformation.

Table 3. We report the number of rejections $R_{0.01}; R_{0.05}; R_{0.20}$ out of 500, using significance levels 0.01, 0.05, and 0.20 in an experiment analogous to the one conducted for Table 1, but now under the parametric null hypothesis H_0^{MEH} with $T = 5{,}000$.

	Transformation-based			Naive			Naive RTC		
Test	KS	CvM	AD	KS	CvM	AD	KS	CvM	AD
N^{MEH}	4; 26; 96	7; 27; 100	10; 31; 106	1; 11; 33	0; 11; 32	3; 14; 40	12; 26; 56	10; 20; 54	30; 72; 96
N^{ExpH}	4; 26; 101	4; 29; 101	4; 30; 94	1; 1; 8	1; 0; 10	1; 1; 12	0; 0; 6	0; 0; 3	2; 7; 26
N^{PLH}	4; 36; 97	5; 30; 99	5; 40; 107	0; 0; 12	0; 0; 7	0; 1; 14	1; 2; 8	1; 2; 7	4; 17; 43

Table 4. We report the number of rejections $R_{0.01}; R_{0.05}; R_{0.20}$ out of 500, using significance levels 0.01, 0.05, and 0.20 in an experiment analogous to Table 3, but where the decay parameters $(\beta_0, \beta_1, \beta_2)$ of the multi-exponential kernel are pre-estimated and kept fixed during likelihood maximization.

	Transformation-based			Naive			Naive RTC		
Test	KS	CvM	AD	KS	CvM	AD	KS	CvM	AD
N^{MEH}	13; 31; 100	13; 25; 95	15; 27; 98	0; 0; 9	0; 0; 7	0; 0; 20	0; 0; 0	0; 0; 0	0; 0; 3
N^{ExpH}	6; 25; 109	3; 26; 105	3; 21; 94	0; 1; 8	0; 0; 8	0; 0; 7	40; 184; 416	20; 201; 449	379; 461; 498
N^{PLH}	6; 28; 107	7; 28; 105	7; 33; 101	0; 0; 9	0; 0; 6	0; 0; 10	0; 0; 17	0; 0; 13	4; 17; 135

6. Application to Real Data

We now consider high-frequency trade data from Société Générale recorded over twelve trading days and restrict attention to the arrival times of trades, which we model as a univariate point process (see Figure 2). The analysis is limited to regular trading hours (9:00–17:30), during which trades arrive on average every three seconds.

To remain consistent with the theoretical framework of Baars et al. [2025], which assumes a constant baseline intensity, each trading day is divided into approximately one-hour windows, within which the baseline is treated as locally constant.

Motivated by the simulation study, we model the data using a multi-exponential Hawkes process, which provides sufficient flexibility to capture multi-scale excitation effects observed in practice. Direct maximum likelihood estimation of all parameters in this model is numerically unstable due to the non-convexity of the likelihood and may lead to pathological estimates, such as branching ratios exceeding one. To obtain a more robust and reproducible fitting procedure, we therefore adopt a two-step estimation strategy.

First, we estimate the decay parameters of the kernel. For each time window, we examine the empirical distribution of inter-arrival times and observe the presence of several distinct temporal regimes. To capture this structure, we model the distribution of the logarithm

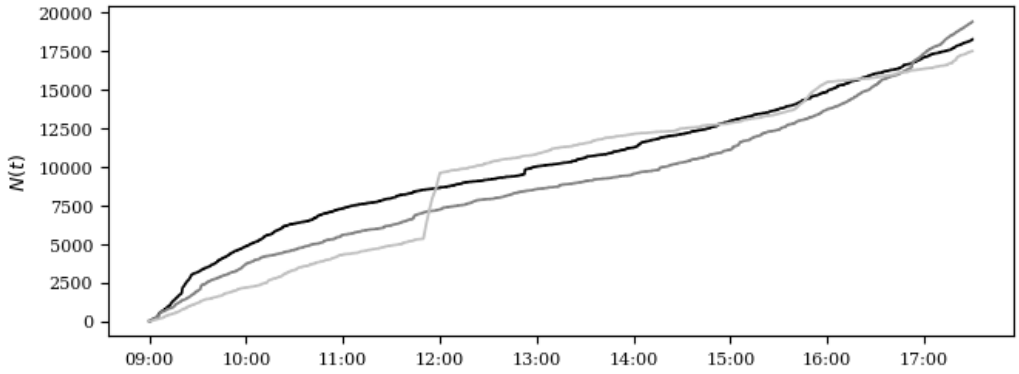

Figure 2. Counting process $N(t)$ for three trading days. From darkest to lightest: Jan. 26, Jan. 25, and Jan. 20.

of inter-arrival times using a Gaussian mixture model (GMM)[1], which provides a flexible approximation of possibly multi-modal distributions.

Formally, a random variable X is said to follow a K-component Gaussian mixture if its density can be written as

$$f_X(x) = \sum_{j=1}^{J} \pi_j \, \mathcal{N}(x \mid \mu_j, \sigma_j^2),$$

where the weights $\pi_j \geq 0$ satisfy $\sum_{j=1}^{J} \pi_j = 1$, and $\mathcal{N}(\mu_j, \sigma_j^2)$ denotes the Gaussian density with mean μ_j and variance σ_j^2. When applied to log inter-arrival times, each Gaussian component can be interpreted as capturing a distinct temporal regime, corresponding to a characteristic time scale of the underlying process.

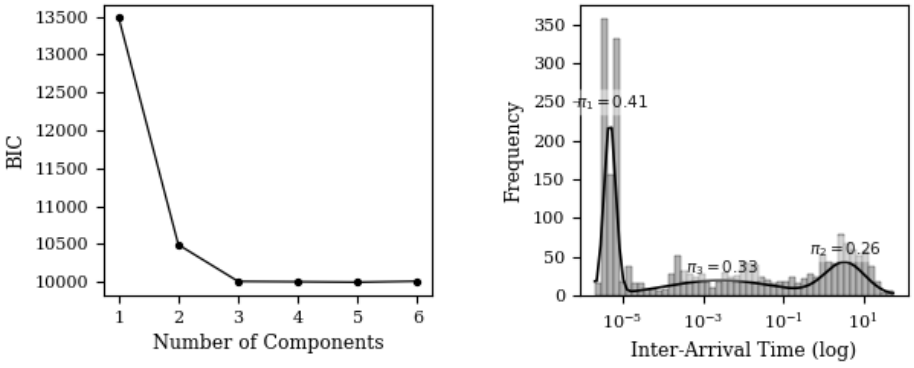

Figure 3. Model selection and inter-arrival time distribution. Left: Bayesian Information Criterion (BIC) as a function of the number of Gaussian mixture components. Right: Empirical distribution of inter-arrival times (histogram, log scale) together with the fitted Gaussian mixture model.

In practice, GMM parameters are estimated by maximum likelihood, and the number of mixture components is selected via the Bayesian Information Criterion (BIC). Across

[1]Fitted using the implementation available in the `scikit-learn` Python library.

the vast majority of time windows, the BIC selects three components, leading us to fix the number of exponential kernels to $J = 3$ in the remainder of the analysis. The corresponding decay parameters are then inferred from the fitted mixture model. Each mixture component is associated with a characteristic time scale τ_j, defined as the median inter-arrival time of that component, $\tau_j = \exp(\mu_j)$, where μ_j denotes the mean of the component on the logarithmic scale. The corresponding decay parameters are set to $\beta_j = 1/\tau_j$.

Second, we assess the stability of the estimated decay parameters across time. We find that these parameters exhibit limited variation across both days and intraday intervals, suggesting that they can be treated as approximately constant as illustrated in Figure 4. We therefore fix the decay parameters to their empirical median values computed across all time windows. Conditional on these fixed decay parameters, we finally estimate the remaining model parameters by maximizing the Hawkes likelihood independently on each time window.

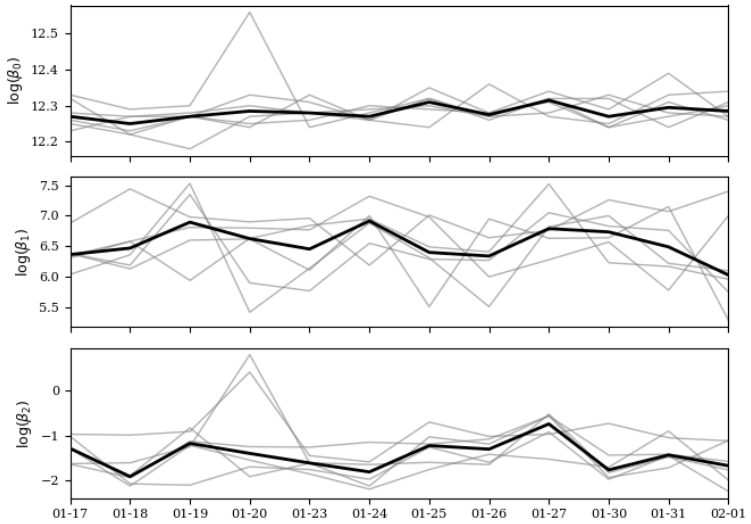

Figure 4. Intraday stability of the estimated decay parameters $\log(\beta_0)$, $\log(\beta_1)$, and $\log(\beta_2)$. Each gray line represents one intraday testing window, while the black line denotes the median across windows for each trading day.

We now report goodness-of-fit test decisions obtained on real data for the exponential, power-law, and multi-exponential Hawkes kernels, with fixed decay parameters in the latter case. For each one-hour testing window, goodness-of-fit is assessed using KS, CvM, and AD tests, and a window is classified as rejected if at least one of the three tests rejects the null hypothesis. Focusing first on the multi-exponential specification, Figure 5 shows that the transformation-based procedure rejects more often than the naive test based on the compensated empirical process, indicating that accounting for parameter estimation uncertainty leads to a more sensitive and less biased assessment of model adequacy. This behavior is consistent with the simulation results and confirms the conservative nature of the naive test on real data.

In contrast, the naive random time change (RTC) procedure rejects almost systematically across time windows. For this reason, and in order to avoid cluttering the presentation

with uninformative results, we do not report the corresponding RTC rejection patterns in Figures 5 and 6. While this behavior clearly signals a lack of perfect fit, it also highlights the extreme sensitivity of the RTC test to parameter estimation uncertainty and to local discrepancies in the fitted intensity, rendering it unreliable as a practical goodness-of-fit diagnostic.

Turning to simpler model specifications, rejection rates increase substantially across all procedures, reflecting stronger model misspecification. In this regime, the naive test may reject even more often than the transformation-based procedure, as the magnitude of model error dominates its conservative bias. Among the two simpler kernels, the power-law specification performs better than the exponential one, likely because its slower decay allows it to capture longer-term dependence effects, although it still falls short of the multi-exponential model.

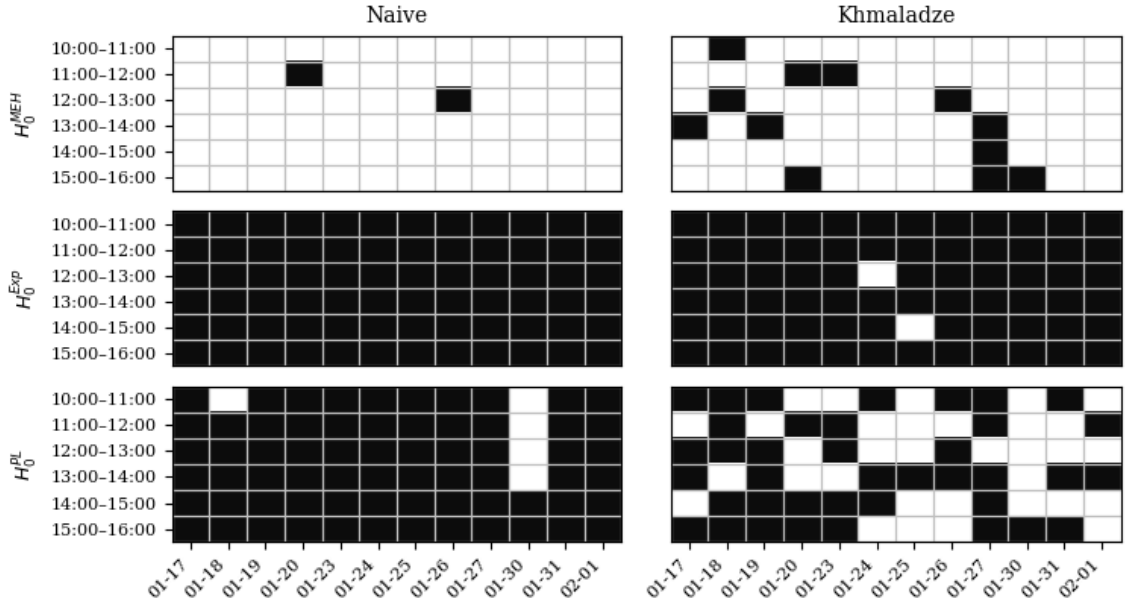

Figure 5. Goodness-of-fit test decisions on real data for three parametric Hawkes model specifications: (top row) the multi-exponential Hawkes model with fixed decay parameters, (middle row) the exponential Hawkes model, and (bottom row) the power-law Hawkes model. Each cell represents a one-hour testing window on a given day and indicates acceptance (white) or rejection (black) of the fitted model. A cell is colored in black if at least one of the three tests (KS, CvM, or AD) rejects the null hypothesis at the 5% significance level; it is white only if none of the three tests rejects.

To further complement this analysis, we investigate the impact of the testing window length on goodness-of-fit decisions for the multi-exponential Hawkes model with fixed decay parameters. Figure 6 reports rejection patterns for window lengths ranging from 15 minutes to 4 hours and reveals a clear monotonic increase in rejection rates as the observation window becomes longer, reflecting the gain in statistical power.

For the naive procedure based on the compensated empirical process, rejection rates remain very low for short windows (1.0% for 15 minutes and 2.1% for 30 minutes), but

rise to 11.1% and 33.3% for 2-hour and 4-hour windows, respectively. The transformation-based procedure exhibits systematically higher rejection frequencies, increasing from 10.4% for 15-minute windows to 69.4% for 4-hour windows. This experiment should not be interpreted as a validation of asymptotic theory, but rather as an empirical illustration of the power properties of the tests. While longer windows provide more informative diagnostics, they also rely on stronger stationarity assumptions, highlighting a fundamental trade-off in practical goodness-of-fit analysis.

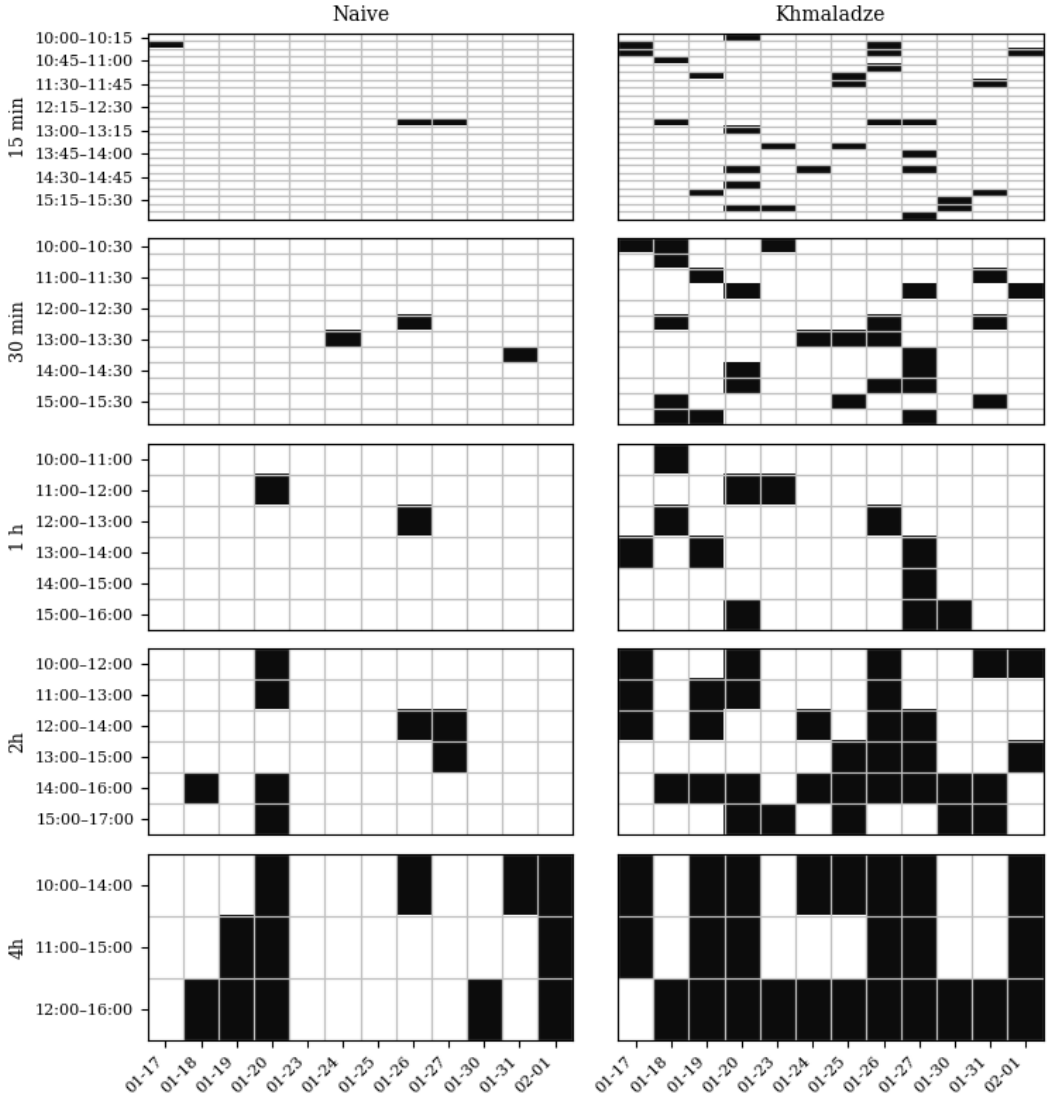

Figure 6. Effect of the testing window length on goodness-of-fit decisions for the multi-exponential Hawkes model with fixed decay parameters. Each cell indicates acceptance (white) or rejection (black) of the fitted model at the 5% significance level.

7. Discussion

Comparison of goodness-of-fit procedures. The empirical results on real data reveal clear and systematic differences between the three goodness-of-fit procedures. For the multi-exponential Hawkes model, the naive test rejects infrequently, in line with its conservative bias under parameter estimation. Accounting explicitly for estimation uncertainty, the transformation-based procedure rejects more often and therefore provides a more sensitive and better calibrated diagnostic of model adequacy. By contrast, the naive random time change (RTC) procedure rejects almost systematically across testing windows, reflecting an extreme sensitivity to parameter estimation uncertainty and to local deviations in the fitted intensity.

These differences become even more pronounced when simpler Hawkes specifications are considered. For the power-law model, the naive test rejects in approximately 93% of the testing windows, compared to 54% for the transformation-based procedure, while the RTC procedure continues to reject almost everywhere. For the exponential model, rejection rates are close to 100% for both the naive and transformation-based procedures, indicating severe misspecification. Together, these results show that when model misspecification is moderate, the transformation-based procedure remains more selective, whereas under strong misspecification, structural model error overwhelms the conservative bias of the naive test.

The Q–Q plots of the RTC-transformed inter-arrival times provide a direct graphical explanation for the systematic rejection of the RTC procedure. As shown in Figure 7, the transformation behaves as expected on simulated data, where empirical quantiles closely follow the theoretical Exp(1) reference line. In contrast, when applied to real data, substantial deviations from exponentiality are observed across the entire range of quantiles, including small and intermediate ones. This indicates that the rejections of the RTC test are not driven solely by tail behavior, but rather reflect a global distortion of the transformed inter-arrival times induced by parameter estimation uncertainty and mild model misspecification.

Limitations. Several limitations of the present analysis should be acknowledged. The transformation-based procedure depends on the choice of hyperparameters, notably the number of increments n and the truncation level τ, which may affect finite-sample performance. The empirical study further relies on the assumption of a locally constant baseline intensity within one-hour testing windows, an approximation that may be violated in practice. Finally, the choice of one-hour windows is somewhat arbitrary and reflects a trade-off between local stationarity and statistical power.

Perspectives. Natural directions for future work include data-driven selection of the hyperparameters (n, τ), adaptive windowing schemes, and extensions to models with time-varying baselines or intraday seasonal components. While the transformation-based goodness-of-fit procedure of Baars et al. [2025] already covers multivariate point processes from a theoretical perspective, further work is needed to assess its practical performance on real multivariate data, where estimation, model complexity, and numerical stability pose additional challenges. Finally, it would be of interest to investigate whether classical random time change–based diagnostics can be modified to explicitly account for parameter estimation uncertainty. As shown by Baars et al. [2025], the bias induced by parameter estimation enters at the same order as the convergence of standard goodness-of-fit statistics and depends on the inter-arrival times themselves, making such corrections an open and

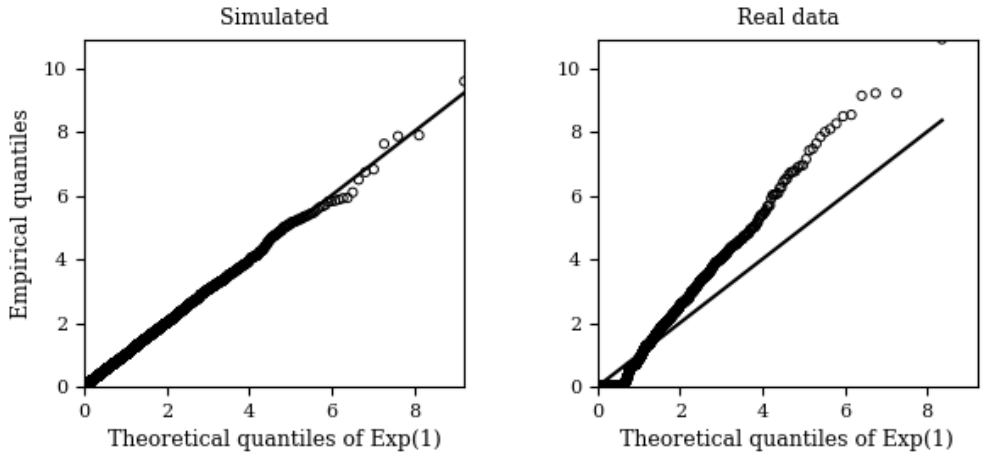

Figure 7. Q–Q plot of the inter-arrival times obtained after applying the random time change transformation on a representative one-hour interval. The solid line corresponds to the theoretical Exp(1) quantiles.

non-trivial problem.

8. Conclusion

In this paper, we reproduced and extended the asymptotically distribution-free goodness-of-fit testing framework proposed by Baars et al. [2025] in the context of univariate Hawkes processes. Through a comprehensive simulation study covering exponential, power-law, and multi-exponential kernels, we confirmed that the transformation-based procedure yields correctly sized tests under the null hypothesis, even in the presence of parameter estimation uncertainty.

Beyond replication, our analysis highlights important practical implications for goodness-of-fit testing in applied point process modeling. Classical procedures that rely on compensators constructed using estimated parameters may lead to misleading conclusions, either by masking model inadequacy or by overstating evidence against a model. In contrast, estimation-aware procedures provide a more reliable diagnostic framework by disentangling genuine model misspecification from artifacts induced by parameter estimation.

From an applied perspective, these results suggest that goodness-of-fit testing for Hawkes processes should not be viewed as a purely mechanical accept–reject decision. Instead, such tests should be used as diagnostic tools, interpreted in conjunction with model complexity, estimation stability, and the intended scope of the application.

Several directions for future research naturally follow from this work, including extensions to multivariate Hawkes models, data-driven selection of tuning parameters, and the development of estimation-aware tests under more flexible semi-parametric or nonparametric specifications.

References

J. Baars, S. U. Can, and R. J. A. Laeven. Asymptotically distribution-free goodness-of-fit testing for point processes. 2025. arXiv:2503.24197.

A. G. Hawkes and D. Oakes. A cluster process representation of a self-exciting process. *Journal of Applied Probability*, 11(3):493–503, 1974.

Y. Ogata. On Lewis' simulation method for point processes. *IEEE Transactions on Information Theory*, 27(1):23–31, 1981.